Explainable Autonomy: A Study of Explanation Styles for Building Clear Mental Models

Explainable Autonomy: A Study of Explanation Styles for Building Clear Mental Models

Garcia, Francisco Javier Chiyah; Robb, David; Liu, Xingkun; Laskov, Atanas ; Patron, Pedro; Hastie, Helen

Publication date:
2019

Document Version
Peer reviewed version

Explainable Autonomy: A Study of Explanation Styles for Building Clear Mental Models

Francisco J. Chiyah Garcia[1], **David A. Robb**[1], **Xingkun Liu**[1],
Atanas Laskov[2], **Pedro Patron**[2], **Helen Hastie**[1]
[1] Heriot-Watt University, Edinburgh, UK
[2] SeeByte Ltd, Edinburgh, UK
{fjc3, d.a.robb, x.liu, h.hastie}@hw.ac.uk
{atanas.laskov, pedro.patron}@seebyte.com

Abstract

As unmanned vehicles become more autonomous, it is important to maintain a high level of transparency regarding their behaviour and how they operate. This is particularly important in remote locations where they cannot be directly observed. Here, we describe a method for generating explanations in natural language of autonomous system behaviour and reasoning. Our method involves deriving an interpretable model of autonomy through having an expert 'speak aloud' and providing various levels of detail based on this model. Through an online evaluation study with operators, we show it is best to generate explanations with multiple possible reasons but tersely worded. This work has implications for designing interfaces for autonomy as well as for explainable AI and operator training.

1 Introduction

Robots and autonomous systems are increasingly being operated remotely in hazardous environments such as in the nuclear or energy sector domains (Hastie et al., 2018; Li et al., 2017; Kwon and Yi, 2012; Nagatani et al., 2013; Shukla and Karki, 2016; Wong et al., 2017). Typically, these remote robots instil less trust than those co-located (Bainbridge et al., 2008; Hastie et al., 2017b; Li, 2015). Thus, the interface between the operator and autonomous systems is key to maintaining situation awareness and understanding between the system and the human operator (Robb et al., 2018). It is this aspect of understanding that we examine here with respect to aligning the operator's mental model (Johnson-Laird, 1980), in terms of both *what* the system can do and *why* it is doing certain behaviours. We propose that this type of explainability will increase trust and therefore adoption of remote autonomous systems.

According to Kulesza et al. (2013), varying the natural language generation of explanations in terms of verbosity (i.e. how many reasons to give or *completeness*) and the level of detail (*soundness*) changes the effectiveness of the explanations in terms of improving the user's mental model. It also affects whether the user thinks that it was "worth it" to read the explanation. It is these aspects of explanation generation that we explore here.

We focus on the natural language generation of explanations as a part of an interactive multimodal system called MIRIAM for situation awareness for autonomous underwater vehicles (AUVs). This interface was developed in conjunction with industry partner SeeByte Ltd (see Figure 1) and runs alongside their commercial UI called SeeTrack with a chat interface, which gives status and mission updates. This multimodal interface has been shown to increase situation awareness (Robb et al., 2018; Hastie et al., 2017a) both by using chat and graphical interface over just graphical interface alone.

We describe a method of explanation generation that is agnostic to the type of autonomy or vehicle. Our contribution is through the 'speak-aloud' method for deriving a model of autonomy for explanations and through the analysis of the forms that these explanations would take to maximally improve the user's mental model. The findings reported here can be used as heuristics for explaining behaviour of remote autonomous systems but also face-to-face robotics (Perera et al., 2016) and other explainable AI tasks such as explaining recommendations (Kulesza et al., 2013). Finally, they could be used to improve operator training.

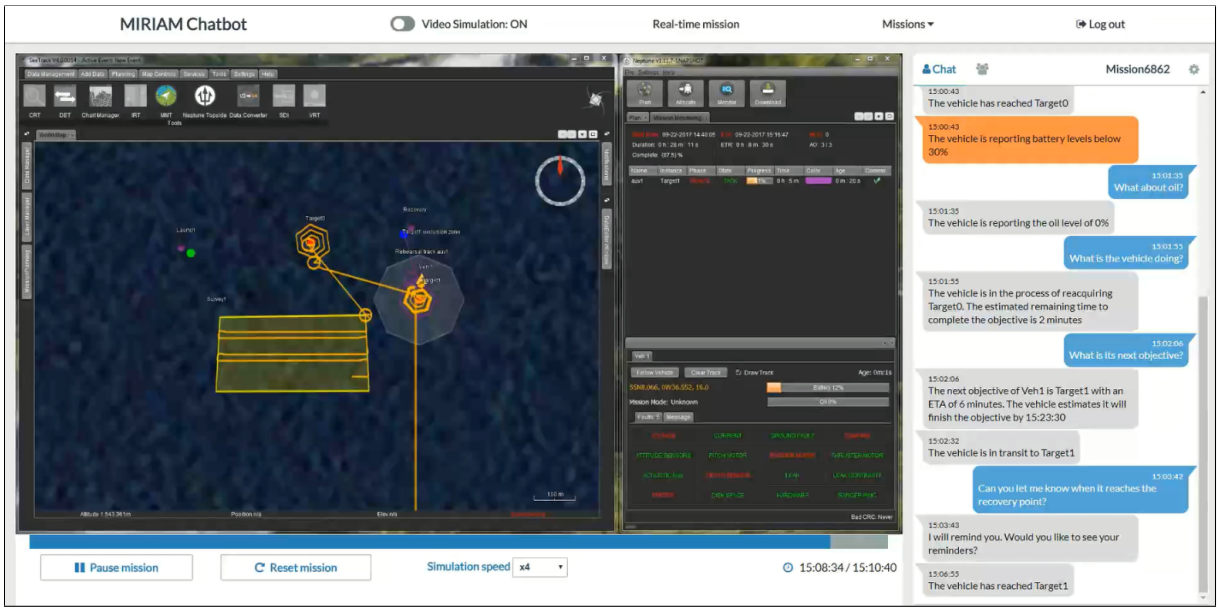

Figure 1: The multimodal interface with SeeTrack interface showing the predicted path of the vehicle on the left and the chat interface on the right where explanations appear.

2 Background

Explainability is an important facet of a transparent system (Wortham et al., 2017) as it can provide the user with a high fidelity mental model, along with increased confidence and performance (Bras et al., 2018; Lim et al., 2009). Mental models, in cognitive theory, provide one view on how humans reason either functionally (understanding what the robot does) or structurally (understanding how it works) (Johnson-Laird, 1980). Mental models are important as they strongly impact how and whether robots and systems are used. In previous work, explainability has been investigated for a variety of systems and users including: 1) explanation of deep learning models for developers, as in (Ribeiro et al., 2016) who showed that such explanations can increase trust; 2) explanations of planning systems (Tintarev and Kutlak, 2014; Chakraborti et al., 2017); and 3) verbalising robot (Rosenthal et al., 2016) or agent (Harrison et al., 2017) rationalisation. Here, we will be looking at verbalising rationalisation of behaviour of the autonomous system, in a similar way to 3). However, these explanations will not be in terms of a constant stream as in (Harrison et al., 2017), rather as part of a mixed-initiative conversational agent where explanations are available on request.

Gregor and Benbasat (1999) describe four types of explanation including "*Why*" and "*Why not*", to explain the functionality and the structure of a system, respectively and *Justification* which includes general knowledge and *Terminological*. Lim et al. (2009) went on to investigate the first two of these and showed that explaining *why* a system behaved a certain way increased both understanding and trust, whilst "*Why not*" showed only an increase in understanding. Here, we will also be investigating these two types of explanations.

We compare our work to that of (Kulesza et al., 2013), who showed that high completeness and high soundness maximised understanding. However, their domain was different to ours (song recommendations) and their users required no specific training or domain knowledge to perform their task. In addition, given the cost of autonomous systems and effort to run missions, the stakes are considerably higher in our case. Adapting explanations to the various users and their existing mental models is touched upon here. Natural language generation has benefited from such personalisation to the user and this applies to explanation generation also. Previous studies in NLG have included adapting to style (Dethlefs et al., 2014), preferences (Walker et al., 2004), knowledge (Janarthanam and Lemon, 2014) and the context (Dethlefs, 2014) of the user. Whilst there has been much work on personalisation of explanations for recommender systems (Tintarev and Masthoff, 2012), there has been little done specifically for explainable AI/Autonomy.

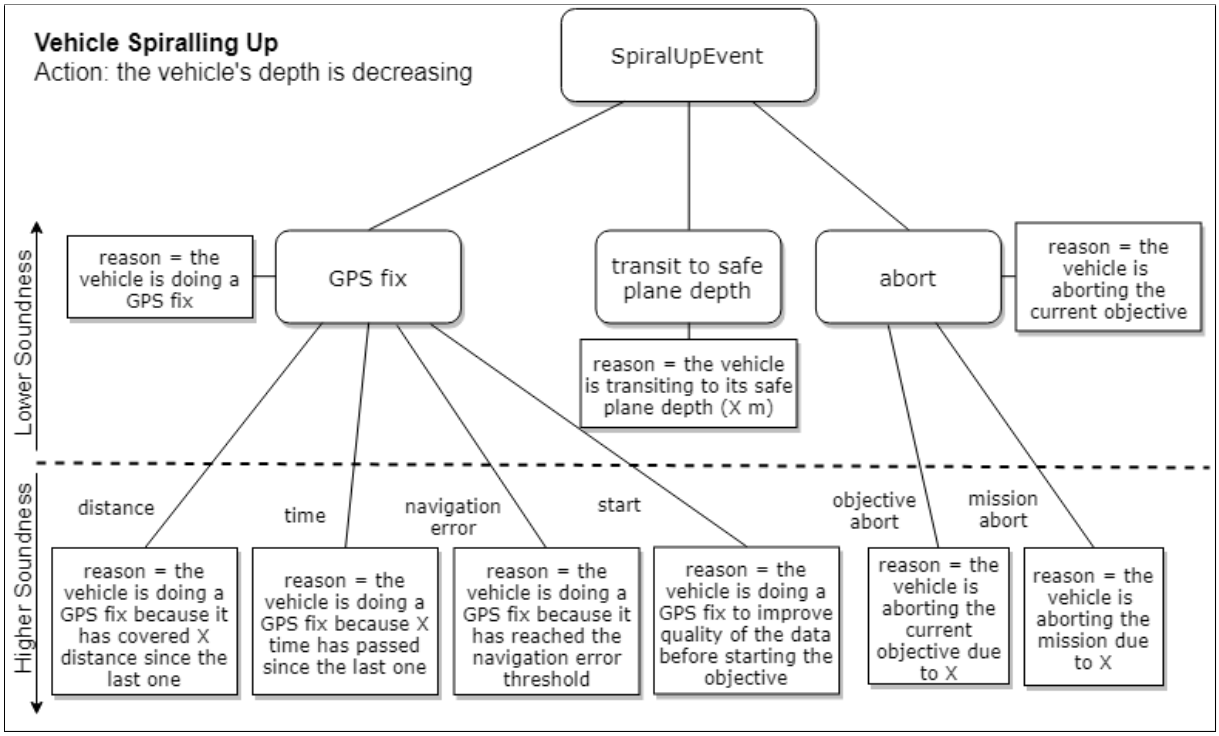

Figure 2: Part of the autonomy model, showing reasons for a vehicle spiralling up. Above/below the dashed line shows what part of the model is used for low/high soundness.

Finally, Gregor and Benbasat (1999) show, users will only take the time to process the explanation if the benefits are perceived to be worth it and do not adversely add to cognitive load (Mercado et al., 2016). Indeed, there needs to be a balance between the amount of information given and the cognitive effort needed to process it. Our evaluation investigates this aspect of explanation generation for our users, who will likely be cognitively loaded given the nature of the task.

3 MIRIAM: The Multimodal Interface

MIRIAM, (Multimodal Intelligent inteRactIon for Autonomous systeMs), as seen in Figure 1, allows for 'on-demand' queries for status and explanations of behaviour. MIRIAM interfaces with the Neptune autonomy software provided by SeeByte Ltd and runs alongside their SeeTrack interface.

MIRIAM uses a rule-based NLP Engine that contextualises and parses the user's input for intent, formalising it as a semantic representation. It is able to process both static and dynamic data, such as names and mission-specific words. For example, it is able to reference dynamic objects such as "auv1", the particular name given to a vehicle in the mission plan, without the requirement to hard-code this name into the system. It can handle anaphoric references over multiple utterances e.g. "Where is Vehicle0?" ... "What is its estimated time to completion?". It also handles ellipsis e.g."What is the battery level of vehicle0?" ..."What about vehicle1?". In this paper, we focus on explanations of behaviours and describe a method that is agnostic to the type of autonomy method. Please refer to (Hastie et al., 2017a) for further details of the MIRIAM system.

4 Method of Explanation Generation

As mentioned above, types of explanations investigated here include *why* (to provide a trace or reasoning) and *why not* (to elaborate on the system's control method or autonomy strategy), a subset of those described in (Gregor and Benbasat, 1999). Lim et al. (2009) show that both these explanations increase understanding and, therefore, are important with regards the user's mental model. We adopt here the 'speak-aloud' method whereby an expert provides rationalisation of the autonomous behaviours while watching videos of missions on the SeeTrack software. This has the advantage of being agnostic to the method of autonomy and could be used to describe rule-based autonomous behaviours but also complex deep learning models. Similar human-provided rationalisation has

been used to generate explanations of deep neural models for game play (Harrison et al., 2017).

The interpretable model of autonomy derived from the expert is partially shown in Figure 2. If a *why* request is made, the decision tree is checked against the current mission status and history and the possible reasons are determined, along with a confidence value based on the information available at that point in the mission[1].

Whilst our explanation generation decides the *content* of the NLG output, the *surface representations* of the explanations are generated using template-based Natural Language Generation (NLG). Templates were picked over statistical surface realisation techniques (e.g. Dethlefs et al. (2014)) due to the fact that the end-user/customer prefers to avoid the variability that comes with statistical methods- these end-users/customers being e.g. the military and operators/technicians in the energy sector. In these domains, vocabulary and standard operating procedures lend themselves to the types of formulaic utterances that template-based systems afford.

The rationalisation of the autonomous behaviours into an intermediate interpretable model, as shown in Figure 2, assists with the uncertainty that remote autonomous systems entail. In our case, communications in the underwater domain are limited and often unreliable. The data received from the vehicles is used to steadily build a knowledge base and generate explanations on-demand. Furthermore, this rationalisation distances the reasoning from the low-level design of the autonomous vehicles to focus on what actually happens during a mission and allows for explanations in broader, high-level terms.

5 Soundness vs Completeness

As mentioned in the Introduction, Kulesza et al. (2013) explore how the level of *soundness* and *completeness* changes how explanations affect the user's mental model, as well as whether the user thinks that it was "worth it" to read the explanation. We adopt Kulesza's terminology here and similarly investigate this trade-off between soundness and completeness. For our domain, an agent that explains the autonomous system using a simpler model reduces soundness (i.e. the top layer of the decision tree, above the line in Figure 2). In this case, the agent provides more general explanations with fewer details that may be easier to digest but may be too broad (see top left of Figure 3).

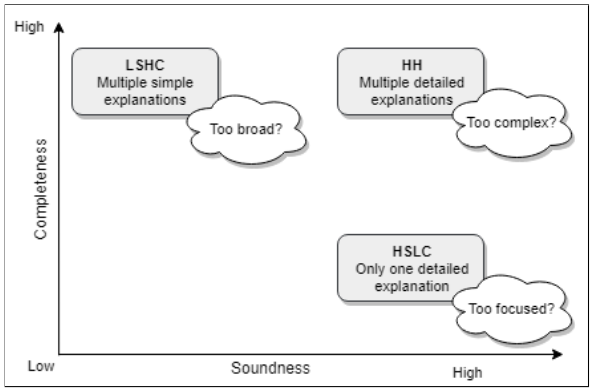

Figure 3: The three types of explanations used in the system, modified from Kulesza et al. (2013): Low Soundness High Completeness (LSHC), High Soundness High Completeness (HH) and High Soundness Low Completeness (HSLC).

High soundness here means that the explanation is taken from the the leaves of the decision tree, thus producing a focused and detailed explanation in Figure 2. An agent with high soundness that gives only one reason, reducing completeness but providing a more concise response, may be viewed as too focused (see bottom right Figure 3)[2]. Combining both high soundness and high completeness may result in too complex an explanation (see top right of Figure 3). We did not include a condition with low soundness and low completeness because it would omit too much data to be relevant or useful in our domain. We investigate these three combinations of varying soundness/completeness and measure their effect on Trust, User Satisfaction and a "worth it" score but primarily the evaluation study focuses on the effect on the user's mental model.

6 Evaluation Method

The experiment was a between-subjects experiment with three conditions, examples of which are given in Table 1. Specifically:

1. C1(HiSoundHiComp): High Soundness, High Completeness - multiple explanations,

[1] above 80% (high), 80% to 40% (medium) and below 40% (low) - levels were determined in consultation with the expert

[2] the one explanation that is presented is the one with the highest confidence at that time -if tied, an ordering that was recommended by the expert is applied

each explaining all of the autonomy model in detail;

2. C2(HiSoundLoComp): High Soundness, Low Completeness - one detailed explanation that explains all of the autonomy model;

3. C3(LoSoundHiComp): Low Soundness, High Completeness - multiple explanations each explaining just the top layer of the autonomy model.

6.1 Experimental Set-up

The experiment consisted of an on-line questionnaire with a pre-questionnaire to gather demographic data and two questions regarding the subjects' pre-existing mental model with regards AUVs: "I have a good understanding of how AUVs work" (Pre-MM-Q1) and "I have a good understanding of what AUVs can do" (Pre-MM-Q2). We were initially looking to investigate trust and so the users were asked to fill out a propensity to trust questionnaire (Rotter, 1967). After the pre-questionnaire, the participants watched 3 scenario videos. After each video, they answered 4 questions regarding the quality of the explanations (US-Q1-4). These questions were modified from the PARADISE-style questionnaire (Walker et al., 1997) for interactive systems and summed to create a User Satisfaction score. In addition, the participants were asked one question on whether the explanations were "worth it" and two questions on their post-explanation mental model (MM-Q1/2). All questions were on a Likert scale with 7 values: from strongly disagree (1) to strongly agree (7).

1. US-Q1: The system chat responses were easy to understand.

2. US-Q2: The system explanations were easy to understand.

3. US-Q3: The system explanations were useful.

4. US-Q4: The system explanations were as expected.

5. "Worth it" question: It would be worth reading the explanations to understand how the system is behaving.

6. MM-Q1: The system explanations in this video help me to increase my understanding of how AUVs work.

7. MM-Q2: The system explanations in this video help me to increase my understanding of what the AUVs were doing.

The mental model questions aim to capture two different dimensions of the user's mental model (Johnson-Laird, 1980): *structurally* so how AUVs work (MM-Q1) and *functionally* so what the AUVs were doing (MM-Q2). We will also refer to the mean of these two scores as the general mental model score, MM-G. After watching the 3 scenarios, a final questionnaire was administrated, which asked about trust and derived a general trust score using the Schaefer scale (Schaefer, 2013).

6.2 The Scenarios

There were three conditions, as discussed above, and for each condition the same three scenario videos were watched by the participants in the same order. Each video showed a mission as reflected in Figure 1 showing the same AUV mission but the text chat running alongside the mission varied depending on the condition. For example, the operator is able to observe in the SeeTrack interface that the vehicle has not done a GPS fix for some time. The operator in the video asks why it is not doing a GPS fix and the answer explains the relevant constraints on the vehicle, as captured in the interpretable autonomy model shown in Figure 2. The different scenarios and the questions are as follows:

1. **Scenario 1:** a *why* question about an event "Why is the vehicle spiralling down?". The system would answer with a few reasons (or only one if completeness was low) with medium confidence. The user would then repeat the question after some time and the system would give a concrete answer with high confidence. See Table 1 for the dialogue for this scenario. Note that in this scenario, the ranking of the most likely reasons does not change during the mission.

2. **Scenario 2:** a *why* question about an event "Why is the vehicle spiralling up?". Similar to Scenario 1 but, in this case, the most likely reason changes in the second answer. This scenario explores how not having all the information (low completeness) affects the user's mental model because the first system's response does not give all possible reasons for the event.

C1 (HiSoundHiComp)	C2 (HiSoundLoComp)	C3 (LoSoundHiComp)
User: Why is the vehicle spiralling up?	**User:** Why is the vehicle spiralling up?	**User:** Why is the vehicle spiralling up?
System: This could be due to the following reasons: - The vehicle is doing a GPS fix at the start of Survey0 to improve the quality of the data gathered (medium confidence) - The vehicle is transiting to its safe plane depth (5m) (medium confidence)	**System:** The vehicle is doing a GPS fix at the start of Survey0 to improve the quality of the data gathered (medium confidence)	**System:** This could be due to the following reasons: - The vehicle is doing a GPS fix (medium confidence) - The vehicle is transiting to its safe plane depth (medium confidence)
System: The vehicle has reached Survey0	**System:** The vehicle has reached Survey0	**System:** The vehicle has reached Survey0
User: Can you confirm now why the vehicle is spiralling up?	**User:** Can you confirm now why the vehicle is spiralling up?	**User:** Can you confirm now why the vehicle is spiralling up?
System: Yes, the vehicle is doing a GPS fix at the start of Survey0 to improve the quality of the data gathered (high confidence)	**System:** Yes, the vehicle is doing a GPS fix at the start of Survey0 to improve the quality of the data gathered (high confidence)	**System:** Yes, the vehicle is doing a GPS fix (high confidence)

Table 1: Example dialogue from Scenario 1 showing varying levels of soundness and completeness for a *why* question.

3. **Scenario 3:** a *why not* question "Why is the vehicle not doing Target0?". The system would compute a reason explaining the procedures of the vehicle in that situation. This scenario investigates how simple explanations (low soundness) about the vehicle's control strategies impact the user's mental model compared to more detailed explanations.

6.3 Participant Group

There were 21 participants (18 male and 3 female[3]). The 21 participants were distributed evenly across the conditions (7 in each). Participation was voluntary and remuneration was by a chance to win one of three £20 Amazon vouchers. The majority of participants were between 25-35 years old, educated to undergraduate, masters degree or PhD level and all worked in the field of software for AUVs, and include roles such as operators and development and software engineers.

For this study, it was important to get users of approximately the same prior mental model of AUVs. Therefore, participants were recruited from a pool of experts in AUVs from industry and academia. This allowed us to design the experiment at a certain level that did not require pre-training of subjects to get to the same expert level. Indeed, the pre-test scores reflect a high self-perceived ability within the participant group with regards their understanding of *how AUVs work* (Pre-MM-Q1 with mean/mode/median/stdev: 6.2/7/6/1) and *what AUVs can do* (Pre-MM-Q2: mean/mode/median/stdev 6.3/6/6/0.6). This approach, however, has the disadvantage of a small pool of users and results in an uneven gender balance. Note that expert levels were evenly spread between conditions.

6.4 Results

Table 2 gives results from the evaluation and shows that C3(LoSoundHiComp) results in higher User Satisfaction scores, "worth it" question and mental model scores. C1(HiSoundHiComp) has the highest level of user trust using the questionnaire from (Schaefer, 2013) with C2 (HiSoundLoComp) having the lowest level of trust, which we discuss below. As indicated in the table, only the mental model questions were found to be statistically significant.

[3] reflecting current gender proportions of employees in the engineering and technology sector, see https://www.theiet.org [accessed May 2018]

	C1 (HiSoundHiComp)		C2 (HiSoundLoComp)		C3 (LoSoundHiComp)	
	Mean **Median**	**SD** **Mode**	**Mean** **Median**	**SD** **Mode**	**Mean** **Median**	**SD** **Mode**
Human-Robot Trust	76.73% 79.29%	6.2% N/A	68.37% 74.29%	13.5% N/A	72.04% 70.00%	13.8% N/A
User Satisfaction	5.56 6	0.695 6	5.51 6	0.615 6	6.06 6	0.693 7
"Worth It" score	5.76 6	0.937 6	5.62 6	0.911 6	6.24 6	0.81 6
MM-Q1 for how work?	5.05 5	1.02 5	4.81 5	1.44 5	5.57* 6	1.66 6
MM-Q2 for what doing?	5.57 6	1.03 6	5.19 5	1.33 5	6.14* 6	1.11 6
MM-G for general MM	5.31 5.5	0.96 5	5 5	1.28 6	5.86* 6	1.23 6

Table 2: Overall descriptive statistics reporting Mean, SD, Median, and Mode. As described in the text, Human-Robot Trust is a score out of 100%. Scales were on a 7 point Likert Scale. User Satisfaction is a scale derived from the average of 4 Likert items. "Worth It" score, MM-Q1 and MM-Q2 are from single Likert scale items. MM-G for general MM is the average of the MM-Q1 and MM-Q2 per participant. N/A for some modes indicates there were no repeated values in that section of the data. We show modes mainly to help describe the sections of the data derived directly from Likert items, i.e. ordinal, but included them across all the data for completeness. These descriptive statistics are for the data aggregated across scenarios within each condition. The * symbols indicate the means of those conditions' distributions which were statistically significantly higher than those of the other two conditions by post hoc Mann-Whitney-U tests following Kruskal-Wallis tests for non-parametric data ($p < .05$) (see text).

Specifically, a Kruskal-Wallis test[4] found a statistically significant effect for these 3 dependant variables across conditions $p < .05$ with χ^2 = 9.3051 for MM-Q1; χ^2 = 9.6836 for MM-Q2, χ^2 = 17.846 for MM-G rejecting the null hypothesis "*there is no difference in the participant's mental model scores between the conditions*". Post-hoc Mann-Whitney-U one-tailed tests using Bonferroni's correction were able to show that C3 was significantly higher than the other two conditions for all three mental models scores at the 95% confidence level. C1 whilst higher than C2 was not significantly so (although there was a trend $p = .02$)[5].

We have also investigated how mental model scores vary across the scenarios during the experiment. We can see from Figure 4 that although C2(HiSoundLoComp) has significantly lower scores than C3(LoSoundHiComp), the user's mental model of how the system works (MM-Q1) builds over time, whereas in conditions C1(HiSoundHiComp) and C3(LoSoundHiComp), it remains steady for the first two scenarios with C2(HiSoundLoComp) actually ending up the highest score by the end of the experiment.

The graph on the bottom of Figure 4 reflects the user's mental model of what the vehicle is doing, which varies from scenario to scenario across conditions. As discussed in Section 6.1, there is a change in confidence in the explanation given in Scenario 2. The system predicts the AUV's action as normal for the first user query, yet in the second query, the system has more data and recomputes the most likely reason, which varies from the one originally presented. Perhaps unsurprisingly, this has a bigger impact on C2(HiSoundLoComp) than on C1(HiSoundHiComp) or C3(LoSoundHiComp) because in those last two conditions, all possible reasons are given so there is less of a surprise compared to the system seemingly 'chang-

[4] A Kruskal-Wallis test was used as MM-Q1/2 are non-parametric and MM-G was shown to be non-normally distributed via a KS Test

[5] $p < .0167$ for significance taking into account Bonferroni's correction

ing its mind' completely. This may also account for the lower general lack of trust for the vehicle in C2(HiSoundLoComp), as indicated in Table 2.

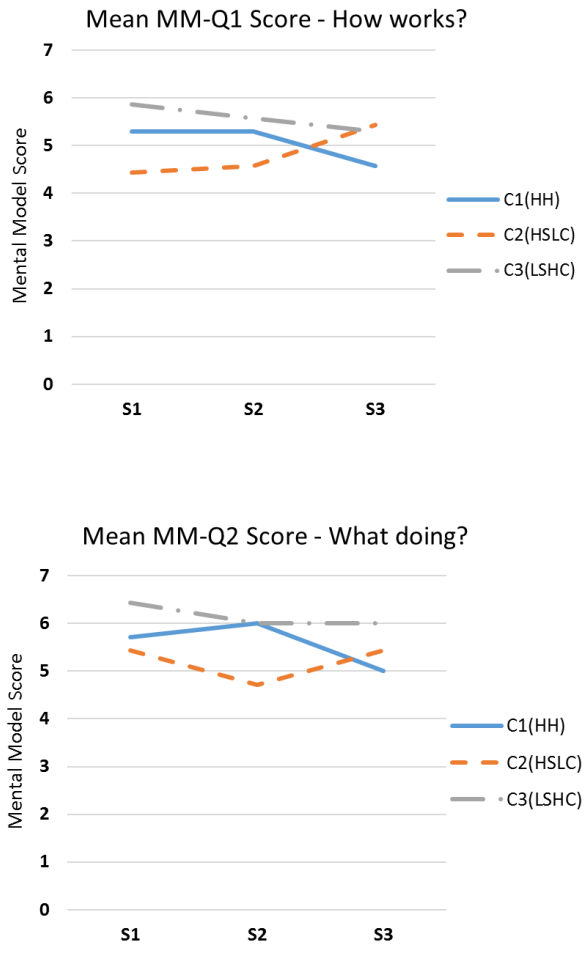

Figure 4: Mean mental model scores across scenarios: C1(HH)–High Soundness High Completeness, C2(HSLC)–High Soundness Low Completeness and C3(LSHC)–Low Soundness High Completeness. S1 to 3: Scenarios 1 to 3.

7 Discussion and Future Work

Kulesza et al. (2013) found that high soundness, high completeness (HiSoundHiComp) explanations performed the best[6]. They found that completeness was linked to better understanding of how the system worked and the highest average mental model scores. They also found that explanations with low completeness resulted in flawed mental models. This is similar to our study where the only condition with low completeness seemed to result in confusion as reflected by significantly lower mental model scores.

In our study, high completeness (i.e. giving all the reasons) is the consistent factor that is important for understanding *how a system works*. However, further investigation is needed to explore the effects of the mental model over longer missions and across missions and to see how the mental models build up in the various conditions, as suggested from Figure 4 where low completeness might be an appropriate presentation method if there is less urgency.

For understanding specific behaviours, i.e. *what the system is doing*, a high level of completeness is important, however a high level of soundness is not necessary (i.e. the reasons don't have to have a lot of detail). In fact, users have a clearer mental model if broader explanations with less details are used with C3(LoSoundHiComp) being statistically higher than the high soundness condition C1(HiSoundHiComp). The difference between our study and that of Kulesza et al. (2013) is that in our study the population have a high degree of pre-existing knowledge and therefore the high soundness may be redundant or even cause frustration or extra cognitive load (Lopes et al., 2018). In addition, according to (Gregor and Benbasat, 1999; Kulesza et al., 2013), "users will not expend effort to find explanations unless the expected benefit outweighs the mental effort". Thus, the system explanations with high soundness, high completeness (HiSoundHiComp) may be too convoluted or distracting in an already complex domain. Our results seem to reflect this trend as well with the "worth it" score, which is highest for C3(LoSoundHiComp). Investigating the cognitive load of processing these various types of explanations is part of future work.

In summary, we present here a method for monitoring and explaining behaviours of remote autonomous systems, which is agnostic to the autonomy model. The positive results from this study suggest that this method produces explanations that build on pre-existing mental models and improves users' understanding of how the systems work and why they are doing certain behaviours. This method, along with recommendations for how explanations should be presented to the user, informs design decisions for interfaces to manage remote autonomous vehicles, as well as explainable autonomy/AI in general.

[6]although no statistical tests were performed due to the low number of subjects

Acknowledgements

This research was funded by EPSRC ORCA Hub (EP/R026173/1, 2017-2021); RAEng/ Leverhulme Trust Senior Research Fellowship Scheme (Hastie/ LTSRF1617/13/37).

References

Wilma A. Bainbridge, Justin Hart, Elizabeth S. Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 701–706, Munich, Germany. IEEE.

Pierre Le Bras, David A. Robb, Thomas S. Methven, Stefano Padilla, and Mike J. Chantler. 2018. Improving user confidence in concept maps: Exploring data driven explanations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM.

Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 156–163, Melbourne, Australia.

Nina Dethlefs. 2014. Context-sensitive natural language generation: From knowledge-driven to data-driven techniques. *Language and Linguistics Compass*, 8(3):99–115.

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser, and Oliver Lemon. 2014. Cluster-based prediction of user ratings for stylistic surface realisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*, pages 702–711, Gothenburg, Sweden. Association for Computational Linguistics.

Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, 23(4):497–530.

Brent Harrison, Upol Ehsan, and Mark O. Riedl. 2017. Rationalization: A neural machine translation approach to generating natural language explanations.

Helen Hastie, Francisco J. Chiyah Garcia, David A. Robb, Pedro Patron, and Atanas Laskov. 2017a. MIRIAM: A multimodal chat-based interface for autonomous systems. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI'17*, pages 495–496, Glasgow, UK. ACM.

Helen Hastie, Xingkun Liu, and Pedro Patron. 2017b. Trust triggers for multimodal command and control interfaces. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI'17*, pages 261–268, Glasgow, UK. ACM.

Helen Hastie, Katrin Solveig Lohan, Mike J. Chantler, David A. Robb, Subramanian Ramamoorthy, Ron Petrick, Sethu Vijayakumar, and David Lane. 2018. The ORCA hub: Explainable offshore robotics through intelligent interfaces. In *Proceedings of Explainable Robotic Systems Workshop, HRI'18*, Chicago, IL, USA.

Srinivasan Janarthanam and Oliver Lemon. 2014. Adaptive generation in dialogue systems using dynamic user modeling. *Comput. Linguist.*, 40(4):883–920.

Philip Nicholas Johnson-Laird. 1980. Mental models in cognitive science. *Cognitive science*, 4(1):71–115.

Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10, San Jose, CA, USA.

Young-Sik Kwon and Byung-Ju Yi. 2012. Design and motion planning of a two-module collaborative indoor pipeline inspection robot. *IEEE Transactions on Robotics*, 28(3):681–696.

Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77:23–37.

Jinke Li, Xinyu Wu, Tiantian Xu, Huiwen Guo, Jianquan Sun, and Qingshi Gao. 2017. A novel inspection robot for nuclear station steam generator secondary side with self-localization. *Robotics and Biomimetics*, 4(1):26.

Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 2119–2129.

José Lopes, Katrin Lohan, and Helen Hastie. 2018. Symptoms of cognitive load in interactions with a dialogue system. In *ICMI Workshop on Modeling Cognitive Processes from Multimodal Data*, Boulder, CO, USA.

Joseph E. Mercado, Michael A. Rupp, Jessie YC. Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in humanagent teaming for Multi-UxV management. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 58(3):401–415.

Keiji Nagatani, Seiga Kiribayashi, Yoshito Okada, Kazuki Otake, Kazuya Yoshida, Satoshi Tadokoro, Takeshi Nishimura, Tomoaki Yoshida, Eiji Koyanagi, and Mineo Fukushima. 2013. Emergency response to the nuclear accident at the fukushima daiichi nuclear power plants using mobile rescue robots. *Journal of Field Robotics*, 30(1):44–63.

Vittorio Perera, Sai P. Selveraj, Stephanie Rosenthal, and Manuela Veloso. 2016. Dynamic generation and refinement of robot verbalization. In *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 212–218, New York, NY, USA. IEEE.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'16, pages 1135–1144, New York, NY, USA. ACM.

David A. Robb, Francisco J. Chiyah Garcia, Atanas Laskov, Xingkun Liu, Pedro Patron, and Helen Hastie. 2018. Keep me in the loop: Increasing operator situation awareness through a conversational multimodal interface. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI'18, Boulder, Colorado, USA. ACM.

Stephanie Rosenthal, Sai P. Selvaraj, and Manuela Veloso. 2016. Verbalization: Narration of Autonomous Robot Experience. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 862–868, New York, NY, USA. AAAI Press.

Julian B. Rotter. 1967. A new scale for the measurement of interpersonal trust. *Journal of Personality*, 35(4):651–665.

Kristin E. Schaefer. 2013. *The Perception and Measurement of Human-Robot Trust*. Ph.D. thesis, College of Sciences at the University of Central Florida, Florida, USA.

Amit Shukla and Hamad Karki. 2016. Application of robotics in onshore oil and gas industry-A review part I. *Robotics and Autonomous Systems*, 75:490–507.

Nava Tintarev and Roman Kutlak. 2014. SAsSy– Making decisions transparent with argumentation and natural language generation. *Proceedings of IUI 2014 Workshop on Interacting with Smart Objects*, pages 1–4.

Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22(4):399–439.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, EACL'97, pages 271–280, Madrid, Spain. Association for Computational Linguistics.

Marilyn A Walker, Stephen J Whittaker, Amanda Stent, Preetam Maloor, Johanna Moore, Michael Johnston, and Gunaranjan Vasireddy. 2004. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840.

Cuebong Wong, Erfu Yang, Xiu-Tian T. Yan, and Dongbing Gu. 2017. An overview of robotics and autonomous systems for harsh environments. In *2017 23rd International Conference on Automation and Computing (ICAC)*, pages 1–6, Huddersfield, UK. IEEE.

Robert H. Wortham, Andreas Theodorou, and Joanna J. Bryson. 2017. Robot transparency: Improving understanding of intelligent behaviour for designers and users. In *Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017*, Lecture Notes in Artificial Intelligence, pages 274–289, Guildford, UK. Springer.